# A Review on Multi-modal Speech Representation Learning

Anonymous ACL submission

## Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style file for ACL 2023. The document itself conforms to its own specifications, and is, therefore, an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

## 1 Introduction

What are speech representations? why would one want to learn them? connections to text and their successes (BERT, ChatGPT)

Motivate speech representation learning then show structure of document. This paper partly references a review on Self-Supervised Speech Representation Learning(Mohamed et al., 2022)

### 1.1 History of Speech Representation Learning

Show first approaches (?)

### 1.2 Self-Supervised Learning for Speech Representation Models

1. Motivation for self supervised learning / What is self supervised learning in general. Where does it maybe originate from?
With the rise of deep learning, the use of labeled data to train models capable of . . . HOW TO INTRO? WITHOUT USING WORDS OF (Mohamed et al., 2022)?

One example of labeled speech data is paired audio-text data, which consists of pairs of voice tracks and corresponding text segments. These kind of data can for example be used to train end-to-end automatic speech representation (ASR) models.
Human-labeled speech data is expensive and generally of limited supply, especially so for languages with comparatively few speakers in the world.
This is why methods using only unlabeled speech data were developed and are since being used to tackle many natural language processing tasks such as the aforementioned ASR task (Kemp and Waibel, 1970).

Self-supervised learning comprises of "techniques that utilize information extracted from the input data itself as the label to learn representations useful for downstream tasks." (Mohamed et al., 2022)
2. Bridge to speech/natural language processing
3. Pre-training. Generative Learning. Contrastive Learning. Predictive Learning.

## 2 Single-mode Speech Representation Models

### 2.1 wav2vec2.0

### 2.2 HuBERT

## 3 Multi-modal Speech Representation Models

This section presents three examples of speech representation learning models each leveraging a different set of input channels. First, we will look at SpeechT5 (Ao et al., 2022) which is using the additional text modality for learning speech representations. Next, we will see AV-HuBERT (Shi et al., 2022), which is an extension of the single-mode model HuBERT () and benefits from the audio modality as well as the video modality. Lastly, VAT-LM () will be covered, leveraging all three mentioned modalities: audio, video and text.

### 3.1 SpeechT5

The SpeechT5 framework, first introduced in 2021 by Microsoft (Ao et al., 2022), is an expansion of the text-only T5 framework (Text-to-Text Transfer Transformer, Raffel et al.'s (2023)). SpeechT5 is. . .

| Model | WER (%) |
|---|---|
| wav2vec2.0 | 1.8 |
| SpeechT5 | 1.5 |

Table 1: ASR Performance.

#### 3.1.1 Model Architecture

#### 3.1.2 Learning (?)

#### 3.1.3 Performance Discussion

Here, we show that SpeechT5 outperforms single-mode SRL models, and also the benefit of using multiple modalities by presenting the ablation studies of the original paper. See Table 1. Also, here is a section of the appendix containing more information: A.

### 3.2 AV-HuBERT

AV-HuBERT Stuffs

#### 3.2.1 Model Architecture

#### 3.2.2 Learning (?)

#### 3.2.3 Performance Discussion

### 3.3 VAT-LM

VAT-LM Stuffs

#### 3.3.1 Model Architecture

#### 3.3.2 Learning (?)

#### 3.3.3 Performance Discussion

## 4 Discussion

Discussion on how well multi-modal models improve the field of SRL.

## Ethics Statement

Scientific work published at ACL 2023 must comply with the ACL Ethics Policy.[1] We encourage all authors to include an explicit ethics statement on the broader impact of the work, or other ethical considerations after the conclusion but before the references. The ethics statement will not count toward the page limit (8 pages for long, 4 pages for short papers).

## Acknowledgements

Maybe acknowledge the review (Mohamed et al., 2022) or smth.

## References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. ArXiv:2110.07205 [cs, eess].

Thomas Kemp and Alex Waibel. 1970. Unsupervised training of a speech recognizer: Recent experiments. *Proc Eurospeech Budapest*, 6.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-Supervised Speech Representation Learning: A Review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210. ArXiv:2205.10643 [cs, eess].

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv:1910.10683 [cs, stat].

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. ArXiv:2201.02184 [cs, eess].

## A Example Appendix

This is a section in the appendix.

[1] https://www.aclweb.org/portal/content/acl-code-ethics